DOCUMENT RESUME

ED 263 177 TM 850 667

AUTHOR Hutchinson, T. P.
TITLE Predicting Performance in Variants of the Multiple-Choice Test.
PUB DATE Jul 85
NOTE 14p.; Paper presented at the Annual Meeting of the Psychometric Society and the Classification Societies (4th, Cambridge, England, July 2-5, 1985).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Guessing (Tests); *Multiple Choice Tests; *Predictive Measurement; Science Tests; Secondary Education; *Statistical Analysis; Test Items; *Test Theory
IDENTIFIERS Answer Until Correct; *Partial Knowledge (Tests)

ABSTRACT

For over 50 years, the overwhelming weight of evidence has been that subjects are able to make use of partial information when responding to multiple-choice items. The subject chooses the alternative which has given rise to the lowest mismatch, except that if this minimum mismatch is larger than some threshold, the question is left unanswered. Assuming some statistical distribution of mismatch, we may obtain the proportions of items answered correctly and answered wrongly, in terms of ability and willingness to guess. This study reanalyzed two datasets. One was an answer-until-correct test of spatial reasoning (386 examinees, 30 items, 5 alternatives). Evidence for the operation of partial knowledge was given by two findings: performance when second and subsequent choices are made (after the first choice is wrong) is above the chance level, and is positively related to first-choice performance. The second dataset was a four-alternative test of chemistry administered to 407 subjects. There were 20 genuine items plus four nonsense items. The proportions of the genuine items answered correctly, answered incorrectly, and omitted can be used to predict the proportion of the nonsense items attempted. Fairly good agreement between predicted and observed proportions was found. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Presented at the Fourth European Meeting of the Psychometric Society and the Classification Societies, held in Cambridge, 2nd-5th July 1985

Predicting performance in variants of the multiple-choice test

T P Hutchinson

Department of Statistics and Operational Research
Coventry (Lanchester) Polytechnic
Priory Street
Coventry CV1 5FB
England

ED263177

Abstract

A possible mathematical description of partial knowledge, ability, and willingness to answer in multiple-choice tests is as follows. Each alternative in each item is envisaged to generate within the subject a certain feeling of mismatch to the question asked. The strength of this tends to be greater for incorrect alternatives than for correct answers, though there is significant random variation. The subject chooses the alternative which has given rise to the lowest mismatch, except that if this minimum mismatch is larger than some threshold, the question is left unanswered. (Of course, the threshold would probably be influenced by the instructions concerning guessing.) Assuming some statistical distribution of mismatch, we may obtain the proportions of items answered correctly and answered wrongly, in terms of ability and willingness to guess.

Many different forms of multiple-choice test have from time to time been used, either for practical or theoretical reasons. Among them have been ones in which: confidence ratings are required; answers are required to items initially omitted; a second choice is permitted when the first answer is wrong; the task is to identify as many as possible of the wrong alternatives; some items are repeated (usually disguised); for some items, no correct answer is among the alternatives available.

The approach outlined in the first paragraph helps us understand performance in such tests. Two datasets have received detailed attention. One was an answer-until-correct test of spatial reasoning (386 examinees, 30 items, 5 alternatives). Evidence for the operation of partial knowledge was given by two findings: (i) performance when second and subsequent choices are made (after the first choice is wrong) is above the chance level, and (ii) is positively related to first-choice performance. The second dataset was a 4-alternative test of chemistry administered to 407 subjects. There were 20 genuine items plus 4 nonsense items. The proportions of the genuine items answered correctly, answered incorrectly, and omitted can be used to predict the proportion of the nonsense items attempted. Fairly good agreement between predicted and observed proportions was found.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
I. P. Hutchinson
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

TM 850 667

BEST COPY AVAILABLE

1. Introduction

Lumsden (1976) began his review of test theory with the comment "there has been a general atmosphere of melancholia and lassitude among latter-day test theorists ... The shreds of theory that have been developed and the time-worn true score models are not rich sources of ideas about testing". And he concluded "The picture revealed is grim ... Little of any consequence has been achieved ... It is only slightly unfair to say that test theory has failed as theory ... I have supped my fill of horrors".

Hutchinson (1982) adapted some ideas from signal detection theory to attempt some advance. He supposed that choosing the correct answer from several alternatives was akin to choosing which of several intervals contained a signal. That is, each alternative gives rise within the subject to a feeling of inappropriateness - inappropriateness, that is, to the question asked. This mismatch between question and answer tends to be higher for the incorrect alternatives than for the correct one, but there is substantial random variation and a significant degree of overlap. Note the analogy with signal detection theory, in which both noise and signal-plus-noise have statistical distributions (which overlap) of their internal representation.

Hutchinson's theory is in many ways crude - there is no attempt at mechanistic realism in modelling the subject's processes of thought in attacking a problem. But at least it is able to predict how the subject will behave when the format of the test is changed - when a second attempt is permitted at questions answered wrongly, for instance. At least the theory is not as bad as assuming the only alternative to perfect knowledge is no knowledge and hence random guessing.

Some empirical evidence was adduced by Hutchinson (1982) in support of his viewpoint:

- When subjects assign a confidence rating to their answer, the higher the level of confidence, the greater is the probability of being correct.
- When an answer is required to items initially left unanswered, a higher-than-chance proportion are found to be correct.
- Scores calculated with the conventional guessing correction (based upon all-or-none knowledge) are higher under "Attempt every item" directions than under "Omit the item if your answer would be a guess" directions.
- When subjects have a second attempt at items answered wrongly, a higher-than-chance proportion are found to be correct.

I have now reanalysed two datasets kindly made available to me: one of responses in an answer-until-correct test, and one of a conventional test with which were included some nonsense items. Brief accounts of the results obtained are given below; further details are given in Hutchinson (1985) for the former, and in Frary and Hutchinson (1982) and Hutchinson (1984) for the latter.

2. A description of partial knowledge analogous to the signal detection model of perception

2.1 Preliminary. Dismissal of models that postulate specific mechanisms

Descriptions of subjects' reactions to some types of item may possess a degree of mechanistic realism. For instance, a subject may know something about a particular alternative answer that eliminates it from consideration. (Asked to indicate whether Paris or Rome is the capital of France, the correct answer is given if the subject knows that Rome is the capital of Italy.) As a second example, the product of $2\frac{1}{2}$ and $3\frac{1}{4}$ may be known to lie between $7\frac{1}{2}$ and 10, thus eliminating alternatives such as $5\frac{1}{4}$ and $11\frac{1}{4}$, without the full details of multiplying fractions being known. I do not quite rule out the feasibility of tailoring theories of partial knowledge to fit specific types of item. But since most tests contain items of many types, and since what is usually wanted is a single score representing some form of general ability, I think a theory of general applicability is to be preferred, even if by its abstractness it loses mechanistic realism.

2.2 Distributions of mismatch

We shall adapt our ideas from signal detection theory (Green and Swets, 1966). To explain errors when a subject is attempting to detect a faint stimulus, this supposes the subject responds according to whether the level of some internal sensation exceeds or falls below a threshold level; and that the sensation is variable (i.e. has some statistical distribution), both when the stimulus is presented and when it is not, the average levels being different in the two conditions. Similarly, we shall suppose that each alternative in each item generates within the subject a certain feeling of inappropriateness to the question posed. This feeling tends to be stronger for the incorrect alternatives than for the correct one, though there is appreciable random variation. The subject normally chooses the alternative that generated the lowest mismatch. But if all exceeded some threshold level, then the subject is unwilling to answer. (This threshold is naturally affected by the instructions given concerning guessing.)

2.3 Mathematical expression

Notation:

N = number of alternatives in each item,

c = proportion of items answered correctly,

w = proportion of items answered wrongly,

u = proportion of items not answered,

X represents the inappropriateness of an alternative. The greater the difference between its average levels for correct and for incorrect alternatives, the easier is the item (or the cleverer is the subject). Denote the distributions of X under the two conditions by F and G:

Probability of X exceeding the value x for correct alternatives = $F(x)$,

Probability of X exceeding x for incorrect alternatives = $G(x)$,

$F(x)$ being less than $G(x)$.

T represents the response threshold, such that if the inappropriateness level exceeds this for all the alternative answers to an item, no choice is made.

We can now write down equations for u, c, and w in terms of F and G:

$$u = F(T)\ \left[G(T)\right]^{N-1}$$

$$c = \int_{-\infty}^{T} \frac{-dF(x)}{dx}\ \left[G(x)\right]^{N-1}\ dx$$

$$w = 1 - u - c$$

What the second of these equations is saying is that the probability of the inappropriateness generated by the correct alternative taking a value x is the probability density corresponding to $F(x)$, $\frac{-dF(x)}{dx}$; that the probability of all the N-1 incorrect alternatives having higher levels of inappropriateness is $\left[G(x)\right]^{N-1}$; that the probability of both these things being true is the product of the probabilities; and, finally, we need to consider all possible values of x less than T, so we sum with T being the limit of integration.

For a given item, ability is measured by how different F and G are. So choose them so as to jointly contain a single parameter characterising ability, λ, and obtain λ in terms of c and w by eliminating T from the above equations. Some examples will show how this is done. (Further implications of this model of performance are described in Hutchinson, 1982.)

2.4 Specific examples

Some choices of F and G give rise to a simple expression for the ability parameter λ.

Firstly, for $0 < x < 1$ and $\alpha \geq 1$, let

$$\left.\begin{aligned} F(x) &= 1 - x \\ G(x) &= \begin{cases} \lambda(1-x)^{1/\alpha} & (1-\lambda^{-\alpha} < x < 1), \\ 1 & (0 < x < 1-\lambda^{-\alpha}). \end{cases} \end{aligned}\right\} \quad (1)$$

In this case, $1-\lambda^{-\alpha} = c-\alpha w/(N-1)$, so that, since the left-hand side of the equation is an increasing function of λ, we have derived the general linear correction for guessing, each correct answer receiving 1 mark and each wrong answer receiving $-\alpha/(N-1)$ marks. The conventional formula is obtained by setting $\alpha = 1$.

Secondly, suppose that for $x > 0$,

$$\left.\begin{aligned} F(x) &= \exp(-x) \\ G(x) &= \exp(-x/\lambda) \end{aligned}\right\} \quad (2)$$

Then $\lambda/(N-1) = c/w$. If $F(x) = 1 - x$ and $G(x) = (1 - x)^{1/\lambda}$, the same equation is obtained, illustrating that a particular formula for λ does not imply a unique pair of functions F and G.

3. Performance in an answer-until-correct test

3.1 Introduction

In answer-until-correct (AUC) tests, the subjects receive immediate feedback as to whether the selected answer was correct or not. If it was wrong, they then choose another answer, and continue attempting the item until the correct answer is found. This practice dates back at least as far as S L Pressey's work in the 1920's. The principal advantages claimed are:

- The greater information per item provided. Hence higher reliability and, perhaps, validity of the test.
- The feedback is liked by the subjects, and the positive attitude produced helps to motivate them.
- Subjects learn from the feedback.

3.2 Theory

The probability of the correct alternative having the second-lowest mismatch is

$$c_2^* = \int_{-\infty}^{\infty} \frac{-dF(x)}{dx} (N-1) [G(x)]^{N-2} [1-G(x)] \, dx$$

The (conditional) probability of giving the correct answer when the second choice is made is thus $c_2 = c_2^*/(1-c)$.

Five choices of pairs of distributions F and G will be considered further. The theories arising will be referrred to as C, ES, E, N, and L: C for a theory that is equivalent to the conventional all-or-nothing learning theory, ES for exponential distributions plus a special state, and E, N, L for exponential, normal, logistic distributions.

C. The simplest case is when F is a rectangular (uniform) distribution over the range 0 to 1 and G is a rectangular distribution over the range $1-\lambda$ to 1 (i e equations (1) with $\alpha=1$). Thus a value of X between 0 and $1-\lambda$ can only arise from the correct alternative - the correct answer is known. Over the range $1-\lambda$ to 1 the ratio of the probability densities is a constant - there is no partial information. c_2 is $1/(N-1)$ whatever c is.

E. The next most straightforward case is when F and G are both exponential distributions over the range 0 to ∞, with different exponents (i e equations (2)). There is no state in which the correct answer is known without error, but a continuous variation of the degree of partial information over the whole range. In the case of 5-alternative items, the relation between c and c_2 turns out to be $c_2 = 4c/(3+c)$.

ES. This case has some features of C and some of E. F is an exponential distribution over the range 0 to ∞, G is an exponential distribution over

the range λ to ∞. A value of X between O and λ can arise only from the correct alternative - the correct answer is known. For larger X, there is a continuous variation of partial information. It is possible to express c as an explicit function of c_2, but not vice versa (see Hutchinson, 1985).

N. F and G are both normal distributions over the range $-\infty$ to ∞, with different means but the same standard deviation. Again, no special state corresponding to certain knowledge. An explicit relation between c and c_2 cannot be obtained, so it is necessary to resort to numerical integration.

L. As N, but with logistic distributions instead of normal distributions.

3.3 Data

The data is that reported by Whetton and Childs (1981). The subjects were 386 school pupils, in the third year of secondary school. The test was designed to give a measure of spatial reasoning ability. It consisted of 30 items all having the same format: each item presents a flag flying on a flagpole around which there is a circle. The flag is then shown in a second picture blowing in a different direction. The subject has to first judge the direction in which the flag is flying relative to the marked position on the circle and then work out where he or she would have to move to on the circle's periphery to see the flag as it is shown in the second picture. Five alternative positions were given, and the test was administered in answer-until-correct format.

3.4 Results

3.4.1 Before resorting to the sophisticated models of 3.2 above, let us observe that certain simple features of the data demonstrate that some form of partial knowledge is operating. Firstly: the proportion of items answered correctly at the second attempt was 0.39 (higher than the chance level of 0.25); the proportions of items answered correctly when it came to the third and fourth attempts were 0.42 and 0.56 (higher than the chance levels of 0.33 and 0.50). Secondly: subjects getting a high proportion of items correct at first attempt tend also to get a high proportion correct at second attempt (correlation = 0.68); subjects getting a high proportion correct within two attempts tend also to get a high proportion correct if a third attempt is necessary (correlation = 0.42).

3.4.2 To compare how well the five theories of 3.2 fit the data, each subject's responses were condensed to three categories - the number of items answered correctly first time, the number for which two attempts

were required, and the number for which three or more attempts were required. Each of the five theories makes a prediction about how these numbers are inter-related, and a value of chi-squared can be calculated to measure the degree to which the data departs from the theory. (Each theory requires one parameter, representing ability, to be fitted to each subject's data.) When the values of chi-squared were summed over all subjects, the following results were found:

C	ES	E	N	L
1770	1269	490	589	565

If a theory were correct, chi-squared would be expected to be 386, since each subject contributes one degree of freedom. So even the best of the theories is not perfect. But more significant is that theory C, implying knowledge is all-or-none, performs much worse than the theories incorporating partial knowledge.

3.4.3 The items varied in difficulty whereas the theories require the estimation of an ability parameter that is the same for all of a number of items for a given subject. Therefore, the analysis was repeated with the test split into three subsets of items, within each of which there was less variation of item difficulty than in the test as a whole. The difficult set consisted of 11 items for which the proportion of subjects answering them correctly at first attempt was between 0.26 and 0.35. The medium set consisted of 11 items for which this proportion was between 0.36 and 0.42. The easy set consisted of 8 items for which this proportion was between 0.45 and 0.71. The results were not appreciably different from those for the whole test.

4. Attempting nonsense items

4.1 Introduction

One means sometimes used to gain more information about the processes operating in multiple-choice tests is to include among the items some which have no correct answer among the alternatives available. This has generally been done for research reasons, not educational ones, though Granich (1931) did suggest announcing their inclusion as a deterrent to random guessing. Inclusion of nonsense items dates back more than 50 years (English, 1928; Thelin and Scott, 1928), but perhaps the largest body of work on this subject is by Slakter and colleagues. He uses the term "risk taking on objective examinations" (rtooe) to refer to the propensity to attempt nonsense items and to Ziller's index (see below) for legitimate items. Slakter (1969) reported the administration to 636 subjects of 4 tests (language aptitude, mathematics aptitude, language achievement, mathematics achievement). These each included 10 nonsense questions embedded in 30 or 40 legitimate questions. Measures of rtooe were calculated from the nonsense items (proportion attempted) and from the legitimate items (by Ziller's method). Slakter found (i) these two measures positively correlated and (ii) rtooe appeared to be a general trait, in the sense that there was a positive correlation between different tests. From this and other studies he concludes that rtooe is a feature of personality, and related to dominance-submission, maladjustment, vocational choice, curriculum choice, and perception of risk in military situations.

4.2 Theory

Following the approach of Section 2, we assume that the probability distribution of mismatch for the alternatives given for the nonsense items is the same as that for the incorrect alternatives in the genuine items. Then the probability of leaving this nonsense item unanswered is $[G(T)]^N$, in which case the probability of giving an answer (denoted a) is $1 - [G(T)]^N$.

If equations (1) hold, then in the special case $\alpha = 1$,

$$a = \frac{Nw}{(N-1)u + Nw} \qquad (3)$$

(it has been assumed that all items are sufficiently difficult for all subjects to have a non-zero probability of giving a wrong answer), or, for general α,

$$a = 1 - \left(\frac{(N-1)u}{(N-1)u + (\alpha+N-1)w} \right)^{N/(\alpha+N-1)} \qquad (4)$$

In the case $\alpha = 1$, equations (1) are equivalent to the conventional model that each subject knows the answer with probability p (specific to the subject, reflecting his or her ability) and decides to guess with probability g if he or she does not, in which case the probability of being correct is 1/N. Then for nonsense items, the probability of response will presumably be g; it turns out that g is given by (3) (Ziller,1957).

Turning now to the case where equations (2) hold, a little algebra shows that

$$a = 1 - u^{Nw/[(N-1)(1-u)]} \qquad (5)$$

4.3 Data

Cross and Frary (1977) report the administration of a 4-alternative 20-item test of chemistry to 407 subjects. As well as the 20 genuine items, there were 4 nonsense items included. The directions to the subjects were designed to encourage informed guessing but discourage wild guessing: "Your score will be the number of items you mark correctly minus a fraction of the number you mark incorrectly. You should answer questions even when you are not sure your answers are correct. This is especially true if you can eliminate one or more choices as incorrect or have a hunch or feeling about which choice is correct. However, it is better to omit an item than to guess wildly among all of the choices given."

4.4 Results

Because each subject was exposed to only 4 nonsense items, the following procedure was adopted. The subjects were grouped into ranges according to their value of expression (3). Then the mean proportion of nonsense items which were answered by the subjects in each group was found for comparison. Finally, the process was repeated with subjects being grouped according to their value of expression (5), rather than (3).

When this was done, it was found that (i) both variants of this theory had some success, in that there was a moderate correlation between the predictions and the findings, (ii) both tended to overestimate, and (iii) formula (5) appeared to be slightly better than formula (3).

Also calculated, this on a subject-by-subject basis, was the correlation between the actual proportion of nonsense items answered (which could only take the values 0, ¼, ½, ¾, 1) and the (two) predicted proportions. This was found to be .46 in the case of expression (3) and .52 in the case of expression (5). Trying different values of α in (4), a maximum correlation of .49 was obtainable: this occurred when $\alpha = 8$. (Though the theories say that a should

take particular values for given u and w, not merely that a should be correlated with them, the possibility that the distractors for the nonsense items may not have been of the same attractiveness as the distractors for the genuine items suggests we should be interested in how high the correlations are, as well as in how small are the differences between predictions and results.)

5. Conclusion

For over 50 years, the overwhelming weight of evidence has been that subjects are able to make use of partial information when responding to multiple-choice items. There has, however, sometimes been a question as to whether they should be permitted to benefit from this. One school of thought says we are trying to estimate the number of test items that the subject knows; and consequently if we have evidence that he or she is getting many right because of partial information, we should make a large deduction from the number he or she gets right in order to obtain the number known; we are therefore penalising partial information. The alternative line of reasoning is that all information is useful, even when it is not complete; that the distinction between full and incomplete information is either not valid or merely a matter of degree; that the subject should be credited for the partial information he or she has. The importance in real life of having to act on incomplete information and make intelligent guesses is adduced in support of this. This dichotomy of opinions has rarely been given explicit attention, though I think it is what Moy and Chou (1982) are getting at in their first paragraph on page five. I take the second view, that partial information is valuable and should be credited, and the structure of the theory of Section 2 reflects this. It is, however, an assumption, and ultimately depends on intuitive notions about the relation between performance on tests and in the real world, and on validity studies of this relation.

I believe I have shown it is practicable (i) to use variant formats of multiple-choice tests to compare different quantitative descriptions of partial knowledge, and (ii) to allow a subject's partial knowledge to contribute to the estimate of his or her ability. There must be many datasets that are potentially suitable for comparing theories: I would be very interested to hear from anyone wanting to collaborate in their analysis.

Acknowledgements

I am very grateful to Prof R B Frary (Virginia Polytechnic Institute and State University, Blacksburg) and to Mr C Whetton (National Foundation for Educational Research, Slough) for making the two datasets available to me.

References

Cross, L.H. & Frary, R.B. (1977). An empirical test of Lord's theoretical results regarding formula scoring of multiple-choice tests. Journal of Educational Measurement, 14, 313-321.

English, H.B. (1928). Bluffing in examinations. American Journal of Psychology, 40, 350.

Frary, R.B. & Hutchinson, T.P. (1982). Willingness to answer multiple-choice questions as manifested both in genuine and in nonsense items. Educational and Psychological Measurement, 42, 815-821.

Granich, L. (1931). A technique for experimentation on guessing in objective tests. Journal of Educational Psychology, 22, 145-156.

Green, D.M. & Swets, J.A. (1966). Signal Detection Theory and Psychophysics. New York: Wiley.

Hutchinson, T.P. (1982). Some theories of performance in multiple choice tests, and their implications for variants of the task. British Journal of Mathematical and Statistical Psychology, 35, 71-89.

Hutchinson, T.P. (1984). Nonsense items in multiple-choice tests. Presented at the London Conference of the British Psychological Society.

Hutchinson, T.P. (1985). Evidence about partial information from an answer-until-correct administration of a test of spatial reasoning. Report from the Department of Statistics and Operational Research, Coventry (Lanchester) Polytechnic.

Lumsden, J. (1976). Test theory. Annual Review of Psychology, 27, 251-280.

Moy, R. & Chou, C.-P. (1982). Interactive computer programs for confidence-marking and answer-until-correct testing. Appendix to B. Choppin, "Latent trait models for answer-until-correct tests", Report from the Center for the Study of Evaluation, Graduate School of Education, University of California, Los Angeles.

Slakter, M.J. (1969). Generality of risk taking on objective examinations. Educational and Psychological Measurement, 29, 115-128.

Thelin, E. & Scott, P.C. (1928). An investigation of bluffing. American Journal of Psychology, 40, 613-619.

Whetton, C. & Childs, R. (1981). The effects of item by item feedback given during an ability test. British Journal of Educational Psychology, 51, 336-346.

Ziller, R.C. (1957). A measure of the gambling response set in objective tests. Psychometrika, 22, 289-292.